4Q-02

日本語電子カルテから ICD コードを自動的に出力する

崎下雅仁† 狩野芳伸†
静岡大学情報学部†

1. はじめに

近年,多くの医療現場における患者の病状の記録媒体が紙から電子的なものに移っている.しかし,世界の医療分野における自然言語処理技術はまだ発達途中にある. 我々は日本語の医療自然言語処理技術の研究の第一歩として,日本語の電子カルテから自動的に診断名を抜き出し,ICD コード1をカルテに割り当てる手法を提案する.ICD コードは世界保健機関 (WHO)によって世界中で病気や死因データ集め,記録,分析をするために作られた.ICD コードは先頭の英字とそれに続く二桁以上の数字で構成される.

実験にはシステムをコンテスト型の国際ワークショップ NTCIR-12 の MedNLPDoc タスク[1]に適用し,このタスクの *Sure* 評価において最も高いパフォーマンスを実現することができた[2].

2. 手法

我々のシステムではカルテの文を入力,ICD コードを出力とする.例えば「肝臓および胸壁に転移あり。」を入力し、この文から ICD コード「C787, C796」を出力する.この節では適切な ICD コードをカルテに付与する 5 つの手法を提案する.手法 2.1 は基礎となる手法であり,手法 2.2〜2.5 は手法 2.1 に加える形で作成した.

2.1. 解析対象文の決定

カルテ中には一文に診断結果を含む文と含まない文がある. 我々のシステムは一文に診断結果を含むとき,その文を解析対象とし,その診断に適切な ICD コードを出力する.

解析対象文か調べるため、参考文献[3]と MedNLPDoc の訓練データの模擬カルテを基に合計 30 (「の診断」など)のキーフレーズを用意した.キーフレーズがあれば解析対象文とするが、文にキーフレーズの否定表現を含んでいたら解析対象文としないこととした.

解析対象文を絞り込んだ後,解析対象文を形態素解析器 Kuromoji2を使い,形態素に分ける.ユーザ辞書には Wikipedia の見出し語に病名を追加したものを用いた.追加した病名については標準病名マスター [4]に書かれている病名を参照した.また,辞書に登録した病名の形態素解析コストを小さくすることで形態素解析結果に病名を出しやすくした.形態素解析結果に病名が現れた場合, 標準病名マスターの表を参照してその病名に相当する ICD コードを割り当てる.

2.2. 医学専門英単語を日本語へ翻訳

カルテの文中には多くの医学専門英単語が現れる.ユーザ辞書に登録されている単語では,英単語が足りず,解決策が必要である.我々はライフサイエンス辞書3を用いて英単語を日本語に翻訳する手法を作成した.翻訳結果は辞書の見出し語に完全一致するものとした.

2.3. 言い換え表現の統一

カルテの文中には多くの言い換え表現が現れる.言い換え表現を統一するために Wikipedia のリダイレクト機能を使うことで解決を試みた.リダイレクト先のある単語はリダイレクト先の単語を用いて統一した.

2.4. 体の部位を含んだ病名の ICD コード出力

2.1 節で説明した手法では,「XX に癌」のような表現が現れた場合,「XX」を無視する.しかし,ここで無視した「XX」は適切な ICD コード出力には必要不可欠な情報である.我々は「悪性新生物」と「損傷」に注目した.

様々な体の部位と「悪性新生物」「損傷」を表す単語の組み合わせから,直接 ICD コードを出力できるルールを作成した.これらのルールは標準病名マスターを参考にして作成した.

一文中に体の部位と悪性新生物を表す単語又は体の部位と損傷を表す単語があればその組み合わせに相当する ICD コードを出力する.損傷の場合,手法 2.1 の解析対象文の絞り込みを行い,悪性新生物の場合はカルテ中全ての文を解析対象文とする.これは悪性新生物には手法 2.1 のキーフレーズに該当しない特別な単語が文中に含まれていることがあるためである.

我々のシステムは「悪性新生物」と「損傷」を表す ICD コードのほぼ全てを出力できる.また手法 2.1 で使用した辞書から「癌」を表す単語と「損傷」を表す単語を取り除いている.それは,それらの単語のみを参考に ICD コードとして出力され,本手法によってカバーできる ICD コードと重複してしまうためである.

Inference of ICD Codes from Japanese Medical Records by Searching Disease Names
Masahito Sakishita† and Yoshinobu Kano†
†Faculty of Informatics, Shizuoka University

1 World Health Organization, International Classification of Diseases (ICD):http://www.who.int/classifications/icd/en/
2 http://www. atilika.org/
3 ライフサイエンス辞書プロジェクト:http://lsd-project.jp/ja/download/

手法の組み合わせ	出力数	完全一致数及びスコア				三桁一致数及びスコア			
		#	P	R	F	#	P	R	F
2.1	424	101	23.82	13.08	16.89	161	37.97	20.85	26.92
2.1+2.2	450	110	24.44	14.25	18.00	176	39.11	22.80	28.81
2.1+2.2+2.3	505	116	22.97	15.03	18.17	185	36.63	23.96	28.97
2.1+2.2+2.3+2.4	575	135	23.48	17.49	20.04	232	40.35	30.05	34.45
2.1+2.2+2.3+2.4+2.5	597	145	24.29	18.78	21.18	245	41.04	31.74	35.79

表 1. 精度 (P), 再現率 (R)、F 値 (F)での手法比較

2.5. XML タグからの ICD コード出力

MedNLPDoc の訓練データは文が XML タグによって分けられている. 中でも既往歴と家族歴タグに注目し,それぞれに特化した ICD コードの出力手法を提案する.既往歴,家族歴は ICD コードが直接的に割り当てられているため選択した.既往歴,家族歴タグに挟まれた文全てを手掛かりとして ICD コードを出力する.2.4 節の手法と同様の方法で,単語と ICD コードの組み合わせを作成した.

3.実験と結果

3.1. 実験

システムの性能を測るため,NTCIR-12 の MedNLPDoc タスクに参加した.MedNLPDoc では訓練データとして 200 のカルテを基にコーパスが提供された.コーパス内のカルテ 1 つあたりの平均文数は 7.82,ICD コード総数は 552 である.テストデータには 78 のカルテが含まれている.正解は,3 人の診断情報管理士4が別々に,同じカルテに ICD コードを付与したものが使われ,また,3 つの評価が設けられ,*Sure* 評価は 3 人のアノテーターが付与した ICD コードのうち全て,*Major* 評価は 2 人以上,*Possible* 評価は一人以上で一致したものある.アノテーター間の一致率は低く,*Sure* 評価が最も信頼できる評価といえる[1].

3.2. 結果

図 1 は *Sure* 評価での全ての MedNLPDoc 参加者の結果を表している.我々の結果は C で,*Sure* 評価の F 値で最も高い値を示している.

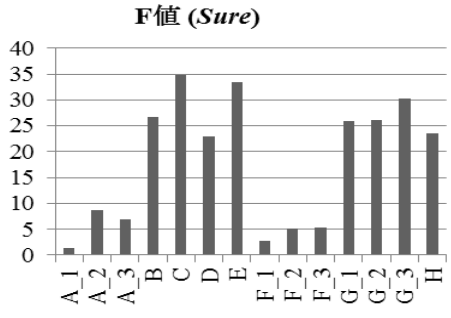

図 1. *Sure* 評価基準における他チームとの F 値比較
(C が我々の結果)

3.2. 各手法の効果分析

各手法の効果は訓練データの結果を用いて実験し,考察する.表 1 は実験結果を示している.「完全一致」は正解 ICD コードと出力との完全一致を,「三桁一致」は正解コードの前半三桁と出力の前半三桁が一致していることを意味する.正解データの合計数は 772 である.2 節で紹介した手法の組み合わせを比較した.

手法 2.2～2.5 のどの手法を手法 2.1 に加えても F 値は上昇するため,どの手法も効果があったといえる.手法 2.4 を加えたとき,F 値の上昇幅が最も大きい.「[体の部位]と損傷」,「[体の部位]と癌」の組み合わせと対応する ICD コードが明確で,ルールが書き易いことが理由であろう.加えて悪性新生物と損傷がカルテに多く表れたことも理由として挙げられる.対して手法 2.3 を加えたとき,F 値の上昇幅は最も小さい.現れた病名や症状を Wikipedia では満足に言い換えられなかったためと考えられる.

4.結論

カルテには病名を含んでいることや言い換え表現があることなどの言語的特徴がある.この特徴から五つの手法を作成することによって ICD コードをより正確に出力できるようになった.提案手法により MedNLPDoc タスクの参加者中で最も高い性能を達成できた.

● 参考文献

[1] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma, “Overview of the NTCIR-12 MedNLPDoc Task,” in proceedings of NTCIR-12 conference, 2016.

[2] M. Sakishita and Y. Kano, “Inference of ICD Codes by Rule-Based Method from Medical Record in NTCIR-12 MedNLPDoc,” in proceedings of NTCIR-12 conference, 2016.

[3] 鳥羽克子, *ICDコーディングトレーニング第2版*. 医学書院, 2006.

[4] K. Hatano and Kazuhiko Ohe, “Information Retrieval System for Japanese Standard Disease-Code Master Using XML Web Service,” American Medical Informatics Association (AMIA) Symposium, 2003.

4 診断情報管理士とは:https://www.jha-e.com/top/abouts/license